Secure Data Transmission and Risk Prediction in Similar Diseases using Convolutional Neural Network

Johncy G
Assistant Professor, Computer Science and Engineering, Anna University, St. Xavier's Catholic college of Engineering, Nagercoil, India

Abstract **- Medical facilities need to be advanced so that better decisions for patient diagnosis and treatment options can be made. Deep learning in healthcare aids the humans to process huge and complex medical datasets and then analyze them into clinical insights. This then can further be used by physicians in providing medical care. Hence deep learning when implemented in healthcare can leads to increased patient satisfaction. In this work, we try to implement functionalities of deep learning in healthcare in a single system. Instead of diagnosis, when a disease prediction is implemented using certain deep learning predictive algorithms then healthcare can be made smart. Some cases can occur when early diagnosis of a disease is not within reach. Hence disease prediction can be effectively implemented. The encrypted health report is uploaded to the cloud server and medical data Provider get the patient report from the cloud server decrypt report by using RSA with DSA key. Then apply Convolutional Neural Network algorithm to find the disease caused by patient based on the symptoms and also find the level of risk of diseases in three stages namely low, medium, high.**

Index Terms **– Deep learning, Disease prediction, CNN, RSA, DSA.**

1.INTRODUCTION

Deep learning is a part of machine learning in artificial Intelligence. It follows the human brain in processing data and developing patterns for use in decision making. It is capable of learning unstructured data. It is also known as deep neural learning or deep neural network[16]. The learning are supervised, semi-supervised or unsupervised. Deep-learning architectures are deep neural networks, deep belief networks, graph neural networks, recurrent neural networks and convolutional neural networks. They are applied in computer vision, machine vision, speech recognition, natural language processing, audio recognition, social network filtering, message translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs. The word “deep” in deep learning describe the use of multiple layers in the network[17]. The deep learning methodologies are classic Neural Networks, Convolutional Neural Networks, Recurrent Neural Network, Generative Adversarial Network, Self Organizing Maps,Boltzmann Machines, Deep Reinforcement Learning, Auto encoders, Backpropagation, Gradient Descent[18].In this paper, We use Convolutional Neural Networks deep learning methodology. CNN is an advanced neural network model. It is made for higher complexity, preprocessing and data compilation[18]. This is also known as shift invariant or space invariant artificial neural networks (SIANN). Convolutional neural network contains three layers. They are input layer, hidden layers and output layer. The hidden layers are convolutional layers, pooling layers, fully connected layers and normalization layers. Convolutional layers get the input and pass its result to the next layer. Pooling layers reduce the size of data by combining the outputs[19]. Fully connected layers arrange the last few layers in the network[20]. Normalization layer is used to arrange the output of the previous layer[21].

2.RELATED WORKS

He et al.[1] described Disease comorbidity method. Disease comorbidity is the presence of one or more diseases along with a primary disorder, which causes additional pain to patients and leads to the failure of standard treatments compared with single diseases. Therefore, the identification of potential comorbidity can help prevent those comorbid diseases when treating a primary disease. By investigating the factors underlying disease comorbidity, e.g. mutated genes and rewired protein-protein interactions (PPIs), they here present a novel algorithm to predict disease comorbidity by integrating multi-scale data ranging from genes to phenotypes. In addition, they identify some pathway and PPI patterns that underlie the co-occurrence between a primary disease and certain disease classes, which can help explain how the comorbidity is initiated from molecular perspectives.

Ni et al.[2] developed a robust and flexible method to integrate tissue-specific molecular networks for disease gene prioritization. This method allows each disease to have its own tissue-specific network(s). When there are multiple tissue-specific networks available for a disease, this method can automatically infer the relative importance of each tissue-specific network. To solve the optimization problem, they develop fast algorithms which have linear time complexities in the number of nodes in the molecular networks. They also provide rigorous theoretical foundations for our algorithms in terms of their optimality and convergence properties.

Kodhai, et al. [3] Data mining is the process of extracting hidden prognostic information from large databases and is a powerful new technology with great potential. However there are mainly two issues namely performance issues and data source issues. Thus an efficient algorithm is proposed to overcome the above issues. The healthcare industry contains large information, which is tedious to process by manual methods. Medical datasets are often not balanced in their class labels. Therefore they proposed an algorithm for healthcare system to accurately predict the result from the large amount of data.

Mastoli, et al.[4] Describe the Machine Learning and Artificial Intelligence has gained much attention from researchers in healthcare and medical sciences. The main purpose of them is to find the best and most suitable algorithm for prediction and diagnosis of diseases and application of machine learning for heathcare systems. They also provides an overview of the data science concepts from data mining technique to machine learning classification algorithms.

Ni, et al. [5] Describe the Quantifying the associations between diseases is now playing an important role in modern biology and medicine. Actually discovering associations between diseases could help us gain deeper insights into pathogenic mechanisms of complex diseases, thus could lead to improvements in disease diagnosis, drug repositioning and drug development. They proposed a new method called ModuleSim to measure associations between diseases by using disease-gene association data and PPIN data based on disease module theory. By considering the interactions between disease modules and their modularity, the disease similarity calculated by ModuleSim has a significant correlation with disease classification of Disease Ontology (DO).

Li, et al. [6] Identifying relatedness among diseases could help deepen understanding for the underlying pathogenic mechanisms of diseases, and facilitate drug repositioning projects. They proposed a new method (MedNetSim) for computing disease similarity by integrating medical literature and protein interaction network. MedNetSim consists of a network-based method (NetSim), which employs the entire protein interaction network, and a MEDLINE-based method (MedSim), which computes disease similarity by mining the biomedical literature. Among function-based methods, NetSim achieved the best performance. They further studied the effectiveness of different data sources.

Qin, et al.[7] Discoveres diseases can provide valuable clues for revealing their pathogenesis and predicting therapeutic drugs. They proposed a framework, namely RADAR, for learning representations for diseases to measure their similarities. RADAR calculates disease similarity by different metrics fully based on the associations between diseases and other disease-related data, and constructs a multi-layer similarity network by integrating multiple disease similarity networks derived from multiple data sources in order to provide a comprehensive evaluation of disease similarities. Experimental results demonstrated that RADAR is effective for detecting similar diseases.

Mathur, et al.[8] Genomics has contributed to a growing collection of gene–function and gene–disease annotations that can be exploited by informatics to study similarity between diseases. They presented functions to measure similarity between terms in an ontology, and between entities annotated with terms drawn from the ontology, based on both co-occurrence and information content. A manually curated dataset with known disease similarities was used as a benchmark to compare the estimation of disease similarity based on gene-based and Gene Ontology (GO) process-based comparisons. GO-Processes associated with similar diseases were found to be significantly regulated in gene expression microarray datasets of related diseases.

Mathur, et.al.[9] The annotation of gene/gene products with information on associated diseases is useful as an aid to clinical diagnosis and drug discovery. They augmented an existing open-disease terminology, the Disease Ontology (DO), and uses it for automated annotation of Swissprot records. Further, they measured disease similarity by exploiting the cooccurrence of annotation among proteins and the hierarchical structure of DO. This makes it possible to find related diseases or signs, with the potential to find previously unknown relationships.

Bandyopadhyay, et al.[10] Describe the Gene Ontology (GO) consists of a controlled vocabulary of terms, annotating a gene or gene product, structured in a directed acyclic graph. Here they introduced a new shortest path based hybrid measure of ontological similarity between two terms which combines both structure of the GO graph and information content of the terms. Here the similarity between two terms t1 and t2, referred to as GOSimPBHMðt1; t2Þ, has two components; one obtained from the common ancestors of t1 and t2. The other from their remaining ancestors. The proposed measure is utilized to compute the average GO similarity score among the genes that are experimentally validated targets of some microRNAs.

Palaniappan, et al.[11] Describe the healthcare industry collects huge amounts of healthcare data which, are not “mined” to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. They developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. IHDPS can answer complex “what if” queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease.

Resnik, et al. [12] presented a new measure of semantic similarity in an is-a taxonomy, based on the notion of information content. Experimental evaluation suggests that the measure performs encouragingly well (a correlation of r = 0 .79 with a benchmark set of human similarity judgments, with an upper bound of r = 0.90 for human subjects performing the same task), and significantly better than the traditional edge counting approach (r = 0 .66).

Amin, et al.[13] Describe Data mining techniques have been widely used in clinical decision support systems for prediction and diagnosis of various diseases with good accuracy. One of the most important applications of such systems is in diagnosis of heart diseases. Heart disease patients have lot of these visible risk factors in common which can be used very effectively for diagnosis. Hence they presented a technique for prediction of heart disease using major risk factors. This technique involves two most successful data mining tools, neural networks and genetic algorithms. The learning is fast, more stable and accurate as compared to back propagation.

Lan, et al.[14] Describe the MicroRNAs (miRNAs) are a type of non-coding RNAs. They proposed a computational framework, KBMF-MDI, to predict the associations between miRNAs and diseases based on their similarities. The sequence and function information of miRNAs are used to measure similarity among miRNAs while the semantic and function information of disease are used to measure similarity among diseases, respectively. In addition, the kernelized Bayesian matrix factorization method is employed to infer potential miRNA-disease associations by integrating these data sources.. The results show that the method can predict unknown miRNA-disease associations.

Zhang, et al. [15] Describe the advances in information technology have witnessed great progress on healthcare technologies in various domains nowadays. They proposed a cyber-physical system for patient-centric healthcare applications and services, called Health-CPS, built on cloud and big data analytics technologies. This system consists of a data collection layer with a unified standard, a data management layer for distributed storage and parallel computing, and a data-oriented service layer. The results of this study show that the technologies of cloud and big data can be used to enhance the performance of the healthcare system.

3.PROPOSED METHODOLOGY

In the previous paper, the system predicts the chronic diseases which are for particular region and for the particular community. The Prediction of Diseases is done only for particular diseases. In the previous paper, Big Data & KNN Algorithm is used for Diseases risk prediction.

In our paper, We combine RSA, DSA algorithm and machine learning algorithm to design a similar diseases prediction and risk prediction system. There are four different entities present in the proposed system namely Admin, Data Owner (patient), Cloud Server (CS) and Medical Data Provider (Doctor). Admin control various different organizations, such as hospitals, medical insurance organizations and medical research institutions. Based on the information submitted by users, admin is responsible for distributing corresponding keys for them. The patient is the owner of their health report, the patient uploads the symptoms along with their health report, which can be encrypted and uploaded to cloud server. Cloud Server is responsible for storing the ciphertext of patient report uploaded by patient. Medical Data Provider get the patient report from the cloud server decrypt report by using decryption key and apply cnn algorithm to find the disease caused by patient based on the symptoms and also find the level of risk of diseases in three stages namely low, medium, high.

3.1.Algorithm

3.1.1. CNN Algorithm

A CNN is a Deep Learning algorithm which can take in an input image, allocate importance to various aspects in the image and be able to differ from one to the other. The pre-processing required in a CNN is much lower as contrast to other classification algorithms. CNN is a subsection of neural network which is inspired by the working principle of using the human visual cortex for object recognition. CNN differs from conventional machine learning algorithms in the context of attribute extraction, where CNN extracts features globally through a number of stacked layers. A CNN is able to successfully capture the Spatial and Temporal dependencies in an image through the application of relevant filters. The architecture performs a better fitting to the image dataset due to the reduction in the number of parameters involved and reusability of weights. In other words, the network can be trained to understand the sophistication of the image better. Generally, CNN architecture consists of several convolution layers and pooling layers. These layers are followed by one or more fully connected (FC) layers. The convolutional layer is the principal building block of a CNN.

In the case of CNNs, the two sets of information are the input data and a convolution filter, which is also called the kernel. The convolutional operation is performed by sliding the kernel over the entire input, which produces a feature map. In order, different filters are utilized to perform multiple convolutions to produce distinct feature maps. These feature maps are finally integrated to formulate the final output from the convolution layer. Activation functions are used after the convolution operation to introduce non-linearity to the model. Different activation functions such as linear function, sigmoid, and tanh are used, but the rectified linear unit (ReLU) was used in the proposed CNN since it can train the model faster and ensure near-global weight optimization. The convolution layer appears before the pooling layer. This layer down-samples each feature map to reduce their dimensions, which in turn reduces over fitting and training time. The max pooling is widely used in CNNs which just selects the maximum value in the pooling window. The Fully Connected layer is basically a fully connected artificial neural network and performs the classification task based on these low-level features. In CNN, the convolution and pooling layers extract low-level features such as edges, lines, ears, eyes, and legs. The activation function used in this final classification layer is typically a SoftMax function.

3.1.2.RSA and DSA Algorithm

RSA (Rivest–Shamir–Adleman) is an algorithm used by modern computers to encrypt and decrypt messages. It is an asymmetric cryptographic algorithm. Asymmetric means that there are two different keys. This is also called public key cryptography, because one of the keys can be given to anyone. The other key must be kept private. The algorithm is based on the fact that finding the factors of a large composite number is difficult: when the factors are prime numbers, the problem is called prime factorization. It is also a key pair (public and private key) generator.

The Digital Signature Algorithm (DSA) appropriate for applications requiring a digital rather than written signature. Digital signature is a pair of large numbers represented in a computer as strings of binary digits. The digital signature is computed using a set of rules and a set of parameters such that the identity of the signatory and integrity of the data can be verified.

RSA involves a public key and private key. The public key can be known to everyone- it is used to encrypt messages. Messages encrypted using the public key can only be decrypted with the private key. RSA is used to transmit shared keys for symmetric key cryptography, which are then used for bulk encryption-decryption. In a public key, the encryption key is public and distinct from the decryption key, which is kept secret (private). An RSA user creates and publishes a public key based on two large prime numbers, along with an auxiliary value. The prime numbers are kept secret. Messages can be encrypted by anyone, via the public key, but can only be decoded by someone who knows the prime numbers. The security of RSA relies on the practical difficulty of factoring the product of two large prime numbers, the "factoring problem". Breaking RSA encryption is known as the RSA problem. Whether it is as difficult as the factoring problem is an open question. There are no published methods to defeat the system if a large enough key is used. RSA is used to transmit shared keys for symmetric key cryptography, which are then used for bulk encryption-decryption.

The DSA provides the capability to generate and verify signatures. Signature generation makes use of a private key to generate a digital signature. Signature verification makes use of a public key which corresponds to, but is not the same as, the private key. Each user possesses a private and public key pair. Public keys are assumed to be known to the public in general. Private keys are never shared. Anyone can verify the signature of a user by employing that user's public key. Signature generation can be performed only by the possessor of the user's private key. Key generation has two phases. The first phase is a choice of algorithm parameters which may be shared between different users of the system, while the second phase computes a single key pair for one user. A hash function is used in the signature generation process to obtain a condensed version of data, called a message digest. The message digest is then input to the DSA to generate the digital signature. The digital signature is sent to the intended verifier along with the signed data (often called the message). The verifier of the message and signature verifies the signature by using the sender's public key. The same hash function must also be used in the verification process. The hash function is specified in a separate standard, the Secure Hash Standard (SHS), FIPS 180. Similar procedures may be used to generate and verify signatures for stored as well as transmitted data.

The DSA is used by a signatory to generate a digital signature on data and by a verifier to verify the authenticity of the signature. Each signatory has a public and private key. The private key is used in the signature generation process and the public key is used in the signature verification process. For both signature generation and verification, the data which is referred to as a message, M, is reduced by means of the Secure Hash Algorithm (SHA) specified in FIPS YY. An adversary, who does not know the private key of the signatory, cannot generate the correct signature of the signatory. In other words, signatures cannot be forged. However, by using the signatory's public key, anyone can verify a correctly signed message. It means of associating public and private key pairs to the corresponding users is required. That is, there must be a binding of a user's identity and the user's public key. This binding may be certified by a mutually trusted party. For example, a certifying authority could sign credentials containing a user's public key and identity to form a certificate. Systems for certifying credentials and distributing certificates are beyond the scope of this standard. NIST intends to publish separate document on certifying credentials and distributing certificates.

4. SYSTEM DESIGN

4.1. Architecture Diagram

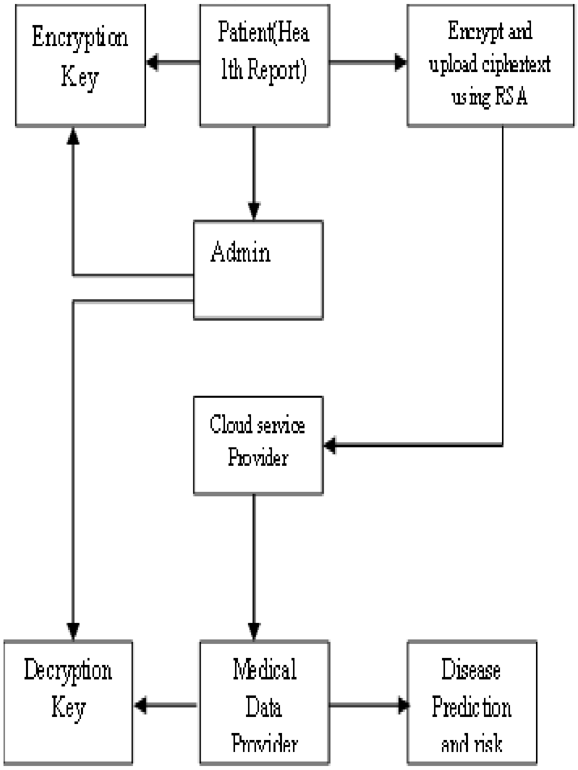


Fig 1. Architecture Diagram

4.2. Data Flow Diagram

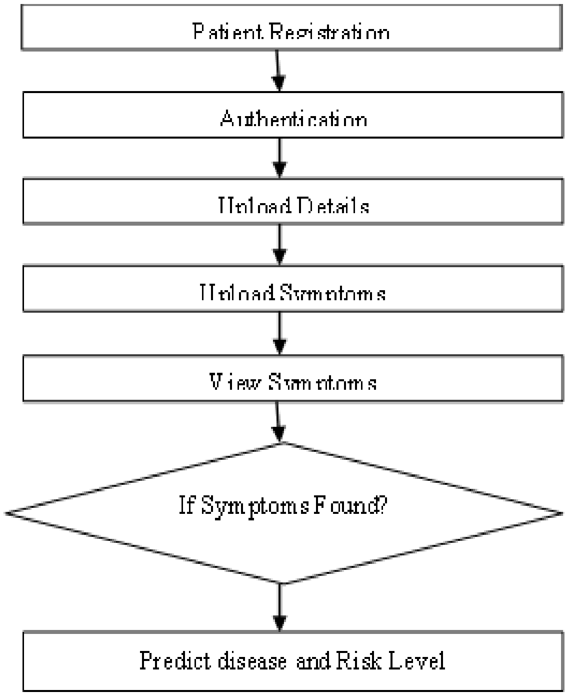


Fig 2 Data Flow Diagram

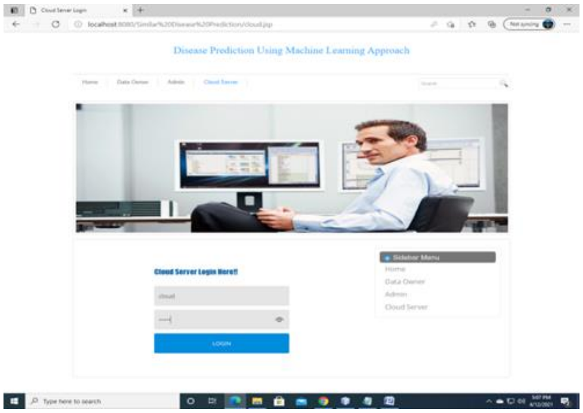
5. RESULTS AND DISCUSSION

At first the user should enter their details along with their symptoms and upload the details in the cloud server. Get the authentication from the cloud server.

After logging in the system as verified admin, the admin can view the patient's detail. Based on the symptoms provided by the patient, the admin will generate the disease report.

This system predicts the disease based on the symptoms provided by the user. It also predicts the risk level of the disease

The system also provides required precautionary measures to overcome a disease.



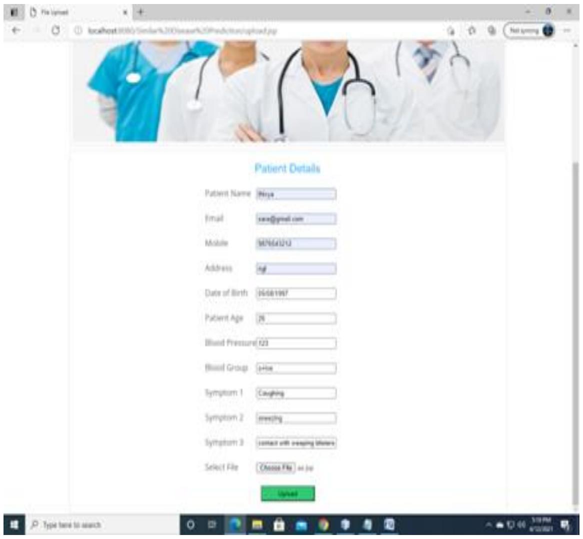
Cloud Server Login Here!!

cloud

LOGIN

Fig 3: cloud server login page

This fig 3 shows the cloud server login page. By logging in the cloud server, user can upload their details in the cloud server.



Patient Details

Patient Name [illegible]

Email sara@gmail.com

Mobile 9876543212

Address hyd

Date of Birth [illegible]

Patient Age 25

Blood Pressure 123

Blood Group o+ive

Symptom 1 Coughing

Symptom 2 sneezing

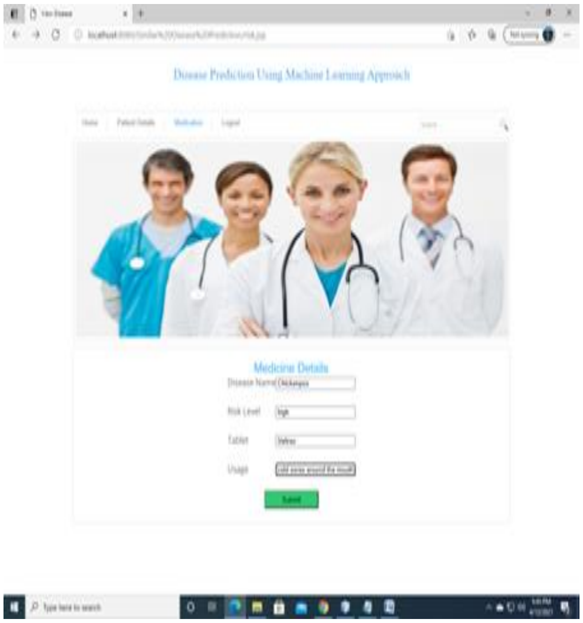
Symptom 3 contact with weeping blisters

Select File Choose File

Upload

Fig 4: Details with the symptoms

In this fig 4 user (patient) provides their details with the symptoms and they will upload a file which consists further clear detail of the patient.



Medicine Details

Disease Name Chickenpox

Risk Level high

Tablet Valtrex

Usage cold sores around the mouth

Save

Fig 5: Predicts the disease

At last this system predicts the disease based on the symptoms provided by the user. It also predicts the risk level of the disease. And it also provides required precautionary measures to overcome the disease.

6. CONCLUSION AND FUTURE ENHANCEMENT

6.1. Conclusion

This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e. predict disease. In conclusion, for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data. We utilized CNN algorithms to classify patient data because today medical data growing very vastly and that needs to process existed data for predicting exact disease based on symptoms. We got accurate general disease risk prediction as output, by giving the input as patients record which help us to understand the level of disease risk prediction. Because of this system may leads in low time consumption and minimal cost possible for disease prediction and risk prediction. The accuracy of our project is 92%.

6.2 Future Enhancement

In future, more disease dataset can be used for classification techniques and other data mining techniques such as clustering can be used to compare the performance of various data mining tools.

REFERENCES

[1] He, Feng, Guanghui Zhu, Yin-Ying Wang, Xing-Ming Zhao, and De-Shuang Huang. "PCID: A novel approach for predicting disease comorbidity by integrating multi-scale data." IEEE/ACM transactions on computational biology and bioinformatics 14, no. 3 (2016): 678-686.

[2] Ni, Jingchao, Mehmet Koyuturk, Hanghang Tong, Jonathan Haines, Rong Xu, and Xiang Zhang. "Disease gene prioritization by integrating tissue-specific molecular networks using a robust multi-network model." BMC bioinformatics 17, no. 1 (2016): 1-13.

[3] Kodhai, E., K. Vagulamaliga, S. Mirudhula, S. R. Pavithra, and K. Rama. "Smart disease prediction using effective vector machine algorithm." International Journal of Pure and Applied Mathematics 116, no. 5 (2017): 55-59.

[4] Mastoli, Ms Manjiri Mahadev, Urmila R. Pol, and Rahul D. Patil. "Machine learning classification algorithms for predictive analysis in healthcare." Mach. Learn 6, no. 12 (2019): 1225-1229.

[5] Ni, Peng, Jianxin Wang, Ping Zhong, Yaohang Li, Fang-Xiang Wu, and Yi Pan. "Constructing disease similarity networks based on disease module theory." IEEE/ACM transactions on computational biology and bioinformatics 17, no. 3 (2018): 906-915.

[6] Li, Ping, Yaling Nie, and Jingkai Yu. "Fusing literature and full network data improves disease similarity computation." Bmc Bioinformatics 17, no. 1 (2016): 1-13.

[7] Qin, Ruiqi, Lei Duan, Huiru Zheng, Jesse Li-Ling, Kaiwen Song, and Xuan Lan. "RADAR: representation learning across disease information networks for similar disease detection." In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 482-487. IEEE, 2018.

[8] Mathur, Sachin, and Deendayal Dinakarpandian. "Finding disease similarity based on implicit semantic similarity." Journal of biomedical informatics 45, no. 2 (2012): 363-371.

[9] Mathur, Sachin, and Deendayal Dinakarpandian. "Automated ontological gene annotation for computing disease similarity." Summit on translational bioinformatics 2010 (2010): 12.

[10] Bandyopadhyay, Sanghamitra, and Koushik Mallick. "A new path based hybrid measure for gene ontology similarity." IEEE/ACM transactions on computational biology and bioinformatics 11, no. 1 (2013): 116-127.

[11] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." In 2008 IEEE/ACS international conference on computer systems and applications, pp. 108-115. IEEE, 2008.

[12] Resnik, Philip. "Using information content to evaluate semantic similarity in a taxonomy." arXiv preprint cmp-lg/9511007 (1995).

[13] Amin, Syed Umar, Kavita Agarwal, and Rizwan Beg. "Genetic neural network based data mining in prediction of heart disease using risk factors." In 2013 IEEE Conference on Information & Communication Technologies, pp. 1227-1231. IEEE, 2013.

[14] Lan, Wei, Jianxin Wang, Min Li, Jin Liu, Fang-Xiang Wu, and Yi Pan. "Predicting microRNA-disease associations based on improved microRNA and disease similarities." IEEE/ACM

transactions on computational biology and bioinformatics 15, no. 6 (2016): 1774-1782.

[15] Zhang, Yin, Meikang Qiu, Chun-Wei Tsai, Mohammad Mehedi Hassan, and Atif Alamri. "Health-CPS: Healthcare cyber-physical system assisted by cloud and big data." IEEE Systems Journal 11, no. 1 (2015): 88-95.